共词网络 LDA 模型的中文文本主题分析:以交通法学文献(2000-2016)为例*

马 红[1] 蔡永明[2]
[1](山东交通学院交通法学院 济南 250357)
[2](济南大学商学院 济南 250022)

摘要:【目的】通过结合传统 LDA 模型的概率主题抽取方法和共词网络分析发现文献词汇间的联系结构的两者优势，降低由少量文献产生的高频词汇的干扰，提高主题凝聚性。【方法】在交通法学文献摘要文本主题分析中，加入文献的关键词作为分词复合词典，提高语义识别度；提出 CA-LDA 模型(Latent Dirichlet Allocation Model with Co-word Analysis)，在传统 LDA 模型的基础上加入共词网络分析，以共词网络拓扑结构参数作为权重控制词汇主题分配(采用介数中心度)，优先提取同时具有高共现性(中介性)和高频率的词汇。【结果】CA-LDA 模型可以得到多篇文献同时共现的高频词汇，这样产生的重点词汇表对主题分析更有意义。该算法的结果不仅仅反映词频概率，同时也能从词汇关联上发现枢纽词汇，更深入理解该领域的研究热点。【局限】CA-LDA 模型主题数目 K 的取值采用混淆度标准交叉验证获得，如果在实际分析中 K 值太大，不利于文献主题的分类整理，未来研究需要对该结果进一步处理来凝聚主题。【结论】本文将该模型应用于交通法学研究领域热点主题分析，在处理大规模文献数据中取得较好效果。相关研究可以拓展应用于各种领域的大规模文献数据自动化处理中。

关键词: 共词网络 LDA 主题模型(CA-LDA) 主题词共现 网络拓扑结构参数 随机梯度下降 交通法学热词
分类号: G254 TP391

1 引 言

信息的不断堆积导致文本的数据量日益庞大。这些文本远远超出一个人的正常阅读能力，同时，越来越多的信息以电子文本的形式存储，为计算机分析文本提供了便利。主题模型(Topic Modeling)能够发现“文档–词语”之间所蕴含的潜在语义关系(即主题)。主题由一个核心事件或活动以及所有与之直接相关的事件和活动组成[1]。利用相关自然语言处理技术，可以对文献内容进行特征分析、提取主题概念、追踪感兴趣的主题，快速、准确获得领域热点知识和发展趋势。主题分析技术已经成为舆情分析、科研选题等方面的有效工具。

主题模型主要采用相似度计算来判断新主题是否属于已知主题，基于统计知识，对文本进行信息过滤，然后利用分类策略跟踪相关主题。目前常用的模型主要有：凝聚层次聚类算法(Hierarchical Clustering Algorithm, HCA)[2-3]，语言模型(Language Model, LM)[4-5]、向量空间模型(Vector Space Model, VSM)[6-7]和概率主题模型(Probabilistic Topic Models, PTM)。其中，潜在狄利克雷分配(Latent Dirichlet Allocation, LDA)模型属于概率主题模型，被公认为是最成功的主题模型。对 LDA 模型的改进主要有快速折叠吉布斯采样 LDA 模型[8]、分布式学习 LDA 模型[9-10]；打破原有可交换的假设的关联

通讯作者：蔡永明，ORCID：0000-0001-7571-1761，E-mail：cymujn@163.com。

*本文系山东省社会科学规划项目“基于复杂网络理论的山东省基础设施系统脆弱性研究”(项目编号：14CGLJ03)、山东省研究生教学创新项目“基于在线学习的研究生学术素养提升开放式生态系统研究”(项目编号：SDYC15045)和济南市哲学社会科学规划项目“济南市网络预约出租车运营状况调查与管理研究”(项目编号: JNSK16C26)的研究成果之一。

LDA 模型[11]; 以及非参数贝叶斯 HDP 模型 (Hierarchical Dirichlet Processes)[12-13]。这些改进极大地提高了主题分析效率, 丰富了 LDA 方法的应用范围。

LDA 模型可以从文本中抽取主题, 但没有考虑多个文本中词汇共现现象。很显然, 词汇在多篇文献中共同出现, 形成的共词网络对于主题凝聚具有指导意义。共词网络分析(Co-word Analysis)是由 Callon 等提出的另一种主题分析技术, 主要分析词汇的共现频率, 通过共词矩阵将距离较近的主题词聚集成簇, 凝聚文献主题[14]。如: Callon 等分析了高分子化学的主题共现情况[15]、Coulter 等研究软件工程主题共现情况[16]、张晓冬等研究计算机集成制造主题共现情况[17], 等等。共词网络分析是一种基于已有主题词的频率及共现的文献关联分析, 并不能产生主题。

因此, 本文结合两者的优势, 提出一种共词网络 LDA 主题模型(CA-LDA), 在传统 LDA 模型中加入共词网络特征参数, 调节主题生成过程。同时, 为了解决新参数带来的计算复杂度, 引入随机梯度下降(Stochastic Gradient Descent, SGD)优化提高了算法执行效率, 在处理大规模文本中取得较好效果。

2 潜在狄利克雷分配模型

潜在狄利克雷分配模型(Latent Dirichlet Allocation, LDA)是由 Blei 等在 2003 年提出的一种概率主题的语言模型[18], 该模型认为任何文本都可以表示成若干潜在主题的混合 Dirichlet 分布, 并可以用词频分布来刻画主题, 以主题混合权重视为 K 维参数的隐含随机变量, 其生成主题的过程如图 1 所示[18-19]。

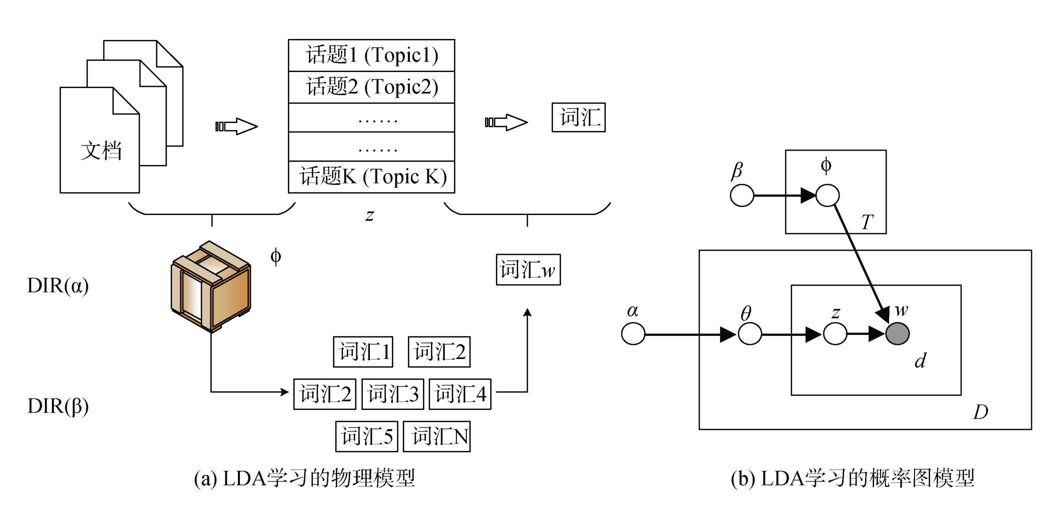


图 1 LDA 学习模型

LDA 的参数估计主要有贝叶斯变分推断(Bayesian Variational Inference, VBI)[20]和 Hoffman 等提出的具有代表性的随机变分方法(Stochastic Variational Inference, SVI)[21]两种方法。传统 LDA 算法中吉布斯采样过程耗时严重, 有时会产生随机梯度噪音, 影响收敛速度。传统 LDA 模型算法过程如下:

(1) 从参数为 α 的 Dirichlet 分布第一取样获得文档主题内容向量 θ, 确定每个主题被选择的概率;

(2) 从主题内容向量 θ 中选择一个主题 z;

(3) 基于一个主题 z 的单词概率分布, 生成单个词汇。

重复此过程, 遍历文档所有词汇, 直到生成所有文档的主题。

主题模型包含语料库 $D=\{W_1, W_2, \cdots, W_M\}$, 文档 d 中的词汇集合 $W=\{w_1, w_2, \cdots, w_N\}$, 所有词汇属于 K 个主题。$z_{dj}$ 代表 d 篇文档的第 j 个单词被划分给主题 z; LDA 的联合概率密度函数[18]为:

$$P(\theta, Z, W \mid \alpha, \beta)=P(\theta \mid \alpha)\prod_{n=1}^{N} P(z_i \mid \theta) P(w_n \mid z_n, \beta) \qquad (1)$$

参数 α 代表文本集上主题的 Dirichlet 分布的先验, 描述了文本集中潜在隐含主题间的相对强弱; β 是一个 K×V 的矩阵, β_{ij} 表示第 i 个主题条件下生成第 j 个单词的概率, 描述了第 j 个特征词归属于第 i 个隐含主题的概率。θ_d 表示文本 d 在 T 个主题上的多项分布, θ 是一个文档级别的主题向量, 每个值对应主题 z 在文

档中出现的概率，z 和 w 都是单词级别的变量，z 由 θ 生成，w 由 z 和 β 共同生成，所有单词 w 分别属于 K 个主题 z。

每一篇文档的潜在主题分布 θ 都服从 Dirichlet 分布，参数 $\alpha_k > 0$ 的情况下(公式(2)) 全部文档集的词频概率[11,18]为公式(3)。

$$P(d|\alpha,\beta)=\int P(\theta|\alpha)(\prod_{n=1}^{N}\sum_{z_n}P(z_n|\theta)P(w_n|z_n,\beta))d\theta \quad (2)$$

$$P(D|\alpha,\beta)=\prod_{d=1}^{M}\int P(\theta|\alpha)(\prod_{n=1}^{N_d}\sum_{z_{dn}}P(z_{dn}|\theta_d)P(w_{dn}|z_{dn},\beta))d\theta_d \quad (3)$$

其中，N_d 代表文档 d 的词汇的总数，对文档中的每一个词 w_n $(1\leqslant n\leqslant N)$，生成一个主题 z_n 服从参数为 θ 的多项式分布。

3　CA-LDA 主题模型

3.1　文本集共词网络构建

共词网络是由文本的主题词在多篇文章或多个段落共同出现(Co-exit)关系构成的一类特殊的科学知识网络。本文研究的摘要文本分析中，共词网络为不同文章摘要中的词汇共现。定义共词网络图 $G(\text{Vertex},\text{Edge})$，其中 Vertex 代表词汇网络节点集合，也即文本集 D 上的全部词汇集 $\text{Vertex}=\{w_1,w_2,\cdots,w_{N_d}\}$； Edge 为词汇共现网络连接的边， $\text{Edge}=\{e_{ij}\,|\,\exists(w_i,w_j),w_i,w_j\in\text{Vertex}\}$，也即词汇 w_i,w_j 在某一文本(或段落)内共现。这样的网络为无向网络，其邻接矩阵 $A_{ij}=\begin{cases}1 & \exists(w_i,w_j),w_i,w_j\in\text{Vertex}\\0 & \text{其他}\end{cases}$，为 $N\times N$ 大规模稀疏矩阵。

复杂网络的拓扑结构特征参数包括：节点连通度指标(如：度 Degree)；中心度指标(如：点度中心度 Degree Centrality、介数中心度 Betweenness Centrality、接近中心度 Closeness Centrality)；节点间紧密度指标(如：簇系数 Clustering Coefficient、派系 Cliques、社区 Community)等。这些参数也表明了一个词汇在共词网络中的重要程度、以及与其他词汇关系的密切程度，可以作为主题生成时计算词汇重要性的参考依据。本文提出的CA-LDA 主题模型使用介数中心度作为词汇归类的调节变量，修正 LDA 模型词汇生成概率，并建立共词网络，提高主题分析的凝聚性(也可以采用点度中心度或接近中心度作为调节变量，其实验结果与介数中心度调节变量的效果基本一致)。

介数中心度，简称中介度，源于社会网络分析中个体的重要性。一个节点的介数中心度表示所有的节点对之间通过该节点的最短路径条数。介数中心度在共词网络中很好地描述了词汇之间的联系的中介关系，以这个词汇为中心的主题归类，可以提高主题内部凝聚性。如果记图中任意两个词汇 w_i,w_j 之间的最短路径条数为 σ_{ij}，而这些最短路径中经过节点 l 的条数为 $\sigma_{ij}(w_l)$，那么节点 w_i,w_j 间经过节点 l 的最短路径条数占 w_i,w_j 间总的最短路径条数的比例为 $\frac{\sigma_{ij}(w_l)}{\sigma_{ij}}$，根据快速介数中心度算法(Faster Algorithm for Betweenness Centrality)[22]，节点 l 的介数中心度定义为：

$$BC(w_l)=\sum_{w_i\in V}\sum_{w_j\neq w_i\in V}\frac{\sigma_{ij}(w_l)}{\sigma_{ij}} \quad (4)$$

传统 LDA 模型给某个文档先选择一个主题 z，再根据该主题生成文档，该文档中的所有词都来自一个主题。主题 $z_1,z_2,\cdots z_K$，生成文档 W 的概率[18]为：

$$P(W)=P(z_1)\prod_{n=1}^{N}P(w_n|z_1)+\cdots+P(z_K)\prod_{n=1}^{N}p(w_n|z_K) \quad (5)$$

CA-LDA 算法的核心是在判断词汇归类时候，考虑词汇在共词网络中的介数中心度。在复杂网络理论中，一个节点的介数中心度越大，该节点在整个网络中就越重要[23]。同理，词汇共现网络 $G(\text{Vertex},\text{Edge})$ 的节点词汇的介数中心度越大，在主题划分中该词汇也越重要。基于这个思想，CA-LDA 模型给生成词汇的概率增加一个权重 $BC(w_1)\Big/\sum_{n=1}^{N}BC(w_1)$ 以控制词汇归类，将传统 LDA 算法生成文档的概率公式(5)修改为公式(6)。这样，介数中心度大的词汇倾向于划分在不同的词袋中，而与该节点词汇关联的词汇倾向于划分到这个主题下。

$$\begin{aligned}P'(W)=&\frac{BC(w_1)}{\sum_{n=1}^{N}BC(w_1)}P(z_1)\prod_{n=1}^{N}P(w_n|z_1)+\cdots\\&+\frac{BC(w_K)}{\sum_{n=1}^{N}BC(w_K)}P(z_K)\prod_{n=1}^{N}P(w_n|z_K)\end{aligned} \quad (6)$$

3.2 随机梯度下降优化

根据吉布斯采样算法[8,18]，对于后验估计 $P(\alpha,\beta|w)=\dfrac{P(\alpha,\beta|w)P(\alpha)P(\beta)}{\iint P(\alpha,\beta|w)d\alpha d\beta}$，如果给定先验初始 α 和 β 相互独立，主题分布 $P(w|\alpha,\beta)$ 可计算出来，通过迭代求出使该式达到最大的 α 和 β。如果增加考虑词向量之间的关联，极大地增加了算法的复杂性。

为了解决这个问题，根据随机梯度下降算法，笔者改进传统吉布斯采样的样本分割与抽样过程，从而降低迭代次数。设计一个随机梯度函数，存储 CA-LDA 模型的参数：主题词汇表 $\{n_k\}_{k=1,v=1}^{K,V}$ 记录词汇 v 分配给主题 k 的频数，词汇表长度为 V，主题数为 K。这样，在每一个吉布斯采样点上，以梯度下降的方向可以最快地获得模型参数 α 和 β。借助吉布斯采样算法[18]的 Gamma 函数 $\gamma_i=\alpha_i+\sum_{n=1}^{N}\phi_{ni}$ 迭代，其中，$\phi_{ni}\propto\beta_i\exp(\Psi(\gamma_i)-\Psi(\sum_{j=1}^{k}\gamma_j))$。文档主题分布的先验参数即可利用梯度下降法求解：

$$\frac{\partial L}{\partial\alpha_i}=M(\Psi(\sum_{j=1}^{K}\alpha_j)-\Psi(\alpha_i))+\sum_{d=1}^{M}(\Psi(\gamma_{dj})-\Psi(\sum_{j=1}^{K}\gamma_j)) \qquad (7)$$

针对每一篇文档的初始 γ 和 ϕ 参数，迭代更新主题词汇表 $\{n_k\}_{k=1,v=1}^{K,V}$，直至收敛即可求出所有主题 z_{ij} 以及最终生成词语 w_{ij}。

3.3 CA-LDA 主题模型算法实现

CA-LDA 算法以及随机梯度下降优化的迭代过程，如图 2 所示。

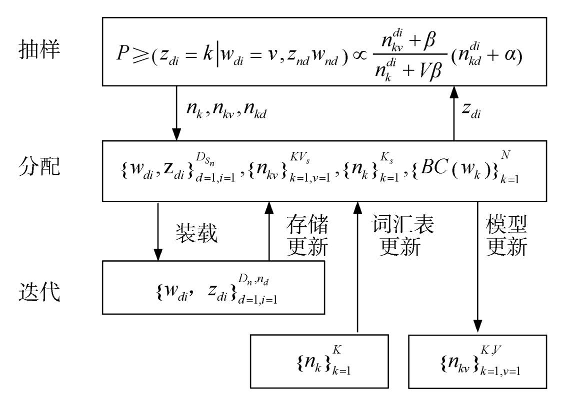


图 2 随机梯度下降主题模型 SGD-LDA 运行过程

具体执行过程可以用伪代码表示：

Initialize $\phi_{n1}^{0}=\frac{1}{k}$ for all i and n

Initialize $\gamma_i=\alpha_i+\frac{N}{k}$ for all i

repeat

  for n=1 **to** N

    for i=1 **to** K

    update $\{w_{di},z_{di}\}_{d=1,i=1}^{D_{S_n}},\{n_{kv}\}_{k=1,v=1}^{KV_s},\{n_k\}_{k=1}^{K_s},\{BC(w_k)\}_{k=1}^{N}$

    $\phi_{ni}^{t+1}=\beta_{iw_n}\exp(\Psi(\gamma_i^t))$

    sample topic: $\{w_{di},z_{di}\}_{d=1,i=1}^{D_n,n_d}$

  end for

  normalize ϕ_{ni}^{t+1} to sum to 1

    $\gamma^{t+1}=\alpha_i+\sum_{n=1}^{N}\phi_n^{t+1}$

until convergence

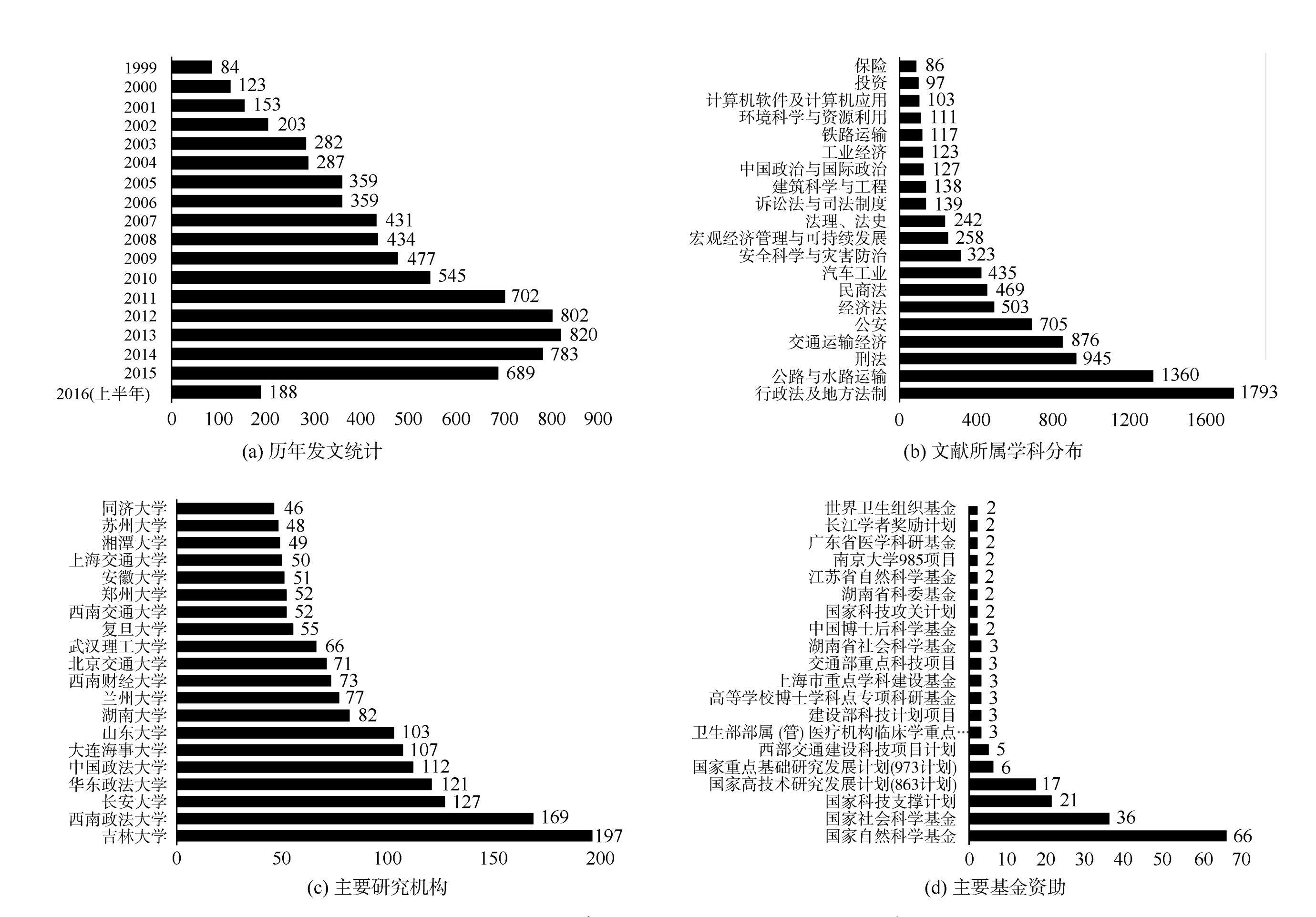
传统 LDA 词汇表来源于概率分布，也就是较高出现频率的词汇作为重点词汇优先提取，而 CA-LDA 模型根据共词网络拓扑结构参数(本文采用中介中心度)调整，获得的结果是同时具有较高的共现性(中介性)和频率的词汇优先提取。这种调整可以降低由少量文献产生高频词汇的干扰，得到多篇文献同时共现的高频词汇，这样产生的重点词汇表对主题分析更有意义。

4 CA-LDA 主题模型交通法学中文文献热点分析

4.1 原始数据获取与描述性统计分析

在 2016 年 7 月 23 日检索中国知网的中国学术期刊网络出版总库，检索式："条件：发表时间 between (2006-01-01, 2016-06-30 and 主题=交通 and 主题=法律 or 主题=法规) (精确匹配)"，检索获得 6 230 条文献记录，根据"发表年份"、"学科"、"机构"和"基金"这 4 项做描述性统计分析，如图 3 所示。

由图 3 中可以看出：交通法学领域研究文献呈现快速增长趋势，但最近两年略有下降；行政法及地方法制、公路与水路运输、刑法、交通运输经济领域的相关文献比较集中；吉林大学、西南政法大学、长安大学、华东政法大学、中国政法大学为主要研究机构；国家自然科学基金、国家社会科学基金、国家科技支撑计划、国家高技术研究发展计划(863 计划)为主要资助来源。



(a) 历年发文统计　　(b) 文献所属学科分布

(c) 主要研究机构　　(d) 主要基金资助

图 3　交通法学研究文献的描述性统计分析

4.2　语料库的生成与信息转化

将交通法学 6 230 篇中文文献的摘要字段提取出来，经过文本整理和分词获得各文档的词汇。采用停用词字典的方法去除文本中部分代词和语气助词等。但如果仅仅做简单分词，得出的高频词汇前 10 位的是：“机制”、“规范”、“建设”、“问题”、“发展”、“管理”、“研究”、“影响”、“社会”、“道路”，这些词汇的内涵不是十分明确，对分析文本主题实际意义并不大。

为此，可以采用增加复合词的方法提高语义识别度，提取 6 230 篇文献的关键词字段，去重后获得 11 565 条词汇作为复合词词典，并将所有复合词分词，一并存储。比对每一篇摘要是否包含该复合词拆分的所有分词，如果包含则去除这些分词，增加该复合词。结果与“关键词”+“摘要”的结果不是一一对应。这样分析的结果可以做到依赖摘要的文本分析而不是作者提供的关键词。

如图 4 所示，以任意一篇文献：中南大学王飞跃发表在《政治与法律》2016 年第 6 期的文章《论道路交通事故责任认定中几对关系的区分》的文本预处理过程为例，简单分词获得 189 个词汇。去除重复词和停用词、增加复合词并删除复合词包含词汇，获得 80 个词汇。其中：“责任推定”、“道路交通事故”、“侵权责任法”、“治安管理”、“刑事责任”、“交通事故”、“交通法律” 7 个词汇为新增复合词。实际上，作者为这篇文章提供的关键词是：“交通事故”、“与交通有关的事故”、“不作为交通违章”、“责任推定” 4 个关键词，两者并不是一一对应。

分析这篇文献新增复合词可以发现：新增复合词有本文关键词，如“责任推定”，该词在其他文献关键词中没有出现过；也有与本文关键词高度相似的复合词，如“道路交通事故”(与本文关键词“交通事故”、“与交通有关的事故”高度相似)来源于2014年北京工业大学孙玉荣发表在《法学杂志》2014 年第 3 期《道路交通事故损害赔偿特殊责任主体研究》，以及湖北警官

论道路交通事故责任认定中几对关系的区分

王飞跃

(中南大学法学院,湖南长沙 410083)

摘要:我国相关法律制度中对交通事故和与交通有关的事故、交通违章中的作为和不作为、责任推定与过错认定等几对关系没有进行准确区分,我国道路交通事故中的责任认定存在诸多问题。用"由车辆引发"、"发生在车辆的运行中"、"由过错或意外引起"等三个因素来限定交通事故,存在过于扩大交通事故"疆域"的弊端,混淆了我国《道路交通安全法》与我国《侵权责任法》、我国《治安管理处罚法》乃至我国《民事诉讼法》、我国《刑事诉讼法》、我国《刑法》等其他法律之间的关系,是导致与交通有关的案件处理错误的主要原因,因而交通事故仍然需要以违章行为为核心进行限定。对交通违章的认识不能仅局限于作为形式,不作为形式的交通违章同样应当受到关注。目前交通事故的责任认定过程,普遍存在以责任推定替代过错认定的错误,导致行政法上的责任转换为民事责任、刑事责任等后果,在交通事故中的责任推定从过错认定中完全剔除之前,至少应当允许不服责任认定的当事人提出反驳,以减少错案的发生。

关键词:交通事故;与交通有关的事故;不作为交通违章;责任推定

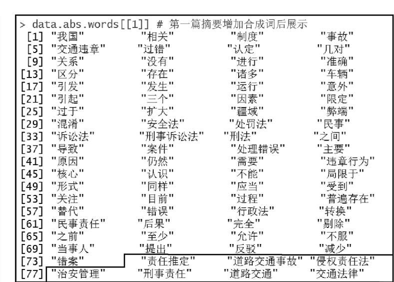
中图分类号:DF31　文献标识码:A　文章编号:1005-9512(2016) 06-0138-07

DOI:10.15984/j.cnki.1005-9512.2016.06.013

(a) 文献摘要原文



(b) 简单分词结果



(c) 预处理后的结果

图 4　语料库的生成中增加复合词处理范例

学院邵祖峰发表在《中国司法鉴定》2012 年第 3 期《论道路交通事故鉴定的现状、问题与对策》等 367 篇文献的关键词; 还有本文关键词没有涉及, 但摘要中出现的复合词, 如"侵权责任法", 则是来源于驻马店市委党校李志浩发表在《广东教育学院学报》2010 年第 6 期《道路交通事故责任主体研究——兼评<侵权责任法>相关规定》等 24 篇文献的关键词。这些复合词的加入与原文摘要内容高度一致, 语义更加明确。

4.3　CA-LDA 主题模型的交通法学热点词汇分析

利用 CA-LDA 模型针对每一篇文章的摘要做主题分析。其中可变量包括超参数 α, ϕ 以及主题数目 K。α 根据主题数目的变化而变化, 由一般经验值可取 $\alpha=\frac{50}{K}$, ϕ 的初始值选 $\phi_{n1}^{0}=\frac{1}{K}$。K 的确定大多采用设置不同的值, 训练后交叉验证(Cross Validation)比较求得最佳值, 其标准一般是采用混淆度(Perplexity) [18], 其计算方法如公式(8)所示, 其中, N_d 为文本 d 的长度(词汇总数); $p(d_d)$ 是待测试模型产生文档 d_d 的概率。混淆度越小, 则模型的泛化能力越强。

$$\text{Perplexity}(D)=\exp\left\{-\frac{\sum_{d=1}^{M}p(d_d)}{\sum_{d=1}^{M}N_d}\right\} \tag{8}$$

根据混淆度计算公式, 在文本集 D 上进行 10 组实验, 获得不同 K 值下的混淆度数值如图 5 所示, 其中 K 值在 50 的时候模型混淆度取得最小值。

利用 CA-LDA 模型, 根据共词网络拓扑结构参数(本文采用中介中心度)调整生成主题概率的权重, 生成 50 个主题。提取各个主题中的前 20 位词汇 1 000 个, 生成共词网络矩阵(1000×1000)稀疏矩阵(Sparsity=98.16%), 权重系数采用 TF-IDF(Term Frequency-Inverse Document Frequency)[18], 去掉稀疏矩阵中低频率的词(Sparsity=90%), 获得 533 个词汇作为领域热点主题词

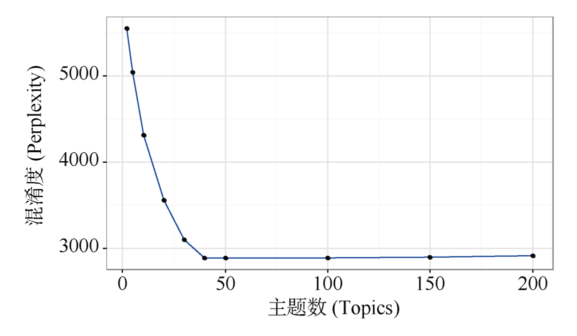


图 5　不同主题数 K 值情况下的混淆度

汇。被去除词汇有"安保公司"、"深水航道"等 467 个词汇, 这些词汇的最高词频为 7, 而剩余词汇词频平均值为 64, 词频最高的"交通安全"达到 353。这种稀疏矩阵降维处理极大减少了计算量, 在大规模文本处理时信息损失较小。

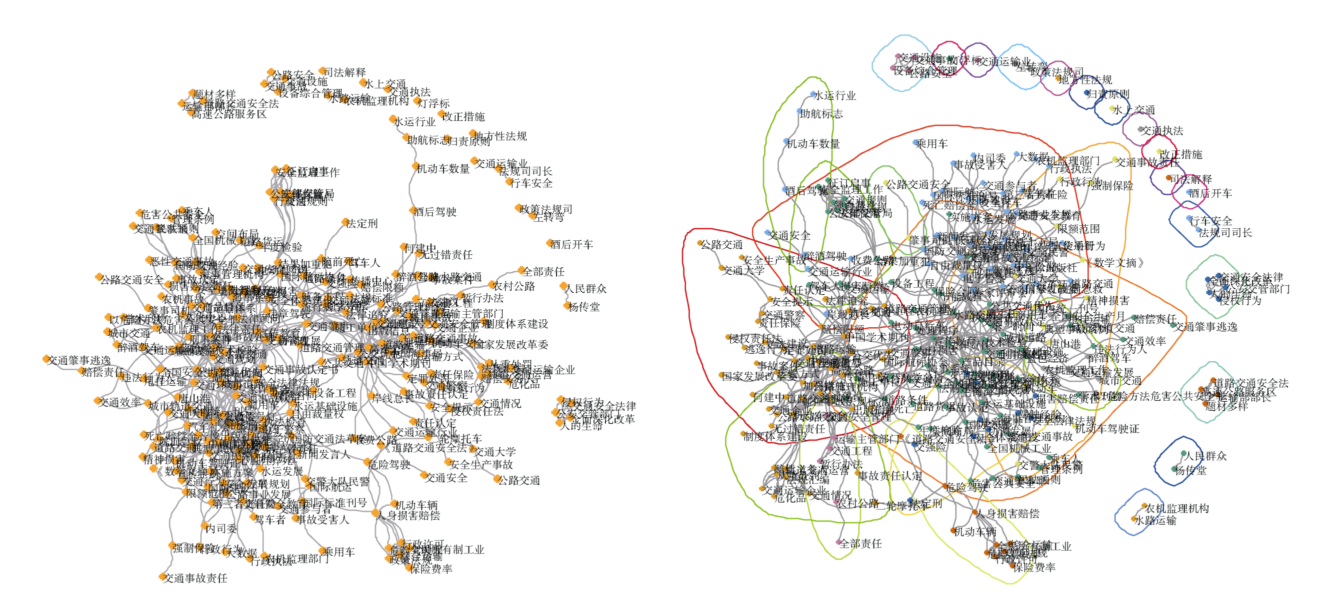
将 CA-LDA 模型获得的 533 个高频主题词汇建立共词网络, 共词网络主题个数降至 28 个, 如图 6 所示。

这些热点词汇基本涵盖 2006 年到 2016 年交通法学研究热点。如果将这些词汇按照出现文献所对应的年份排序, 可以发现热点领域演变。

4.4　CA-LDA 模型方法与传统 LDA 模型主题分析结果比较

在同一数据集上(交通法学 6 230 篇中文文献的摘要文本), 分别采用 CA-LDA 模型方法与传统 LDA 模型主题做分析实验, 结果如表 1 所示。由于 CA-LDA 和传统 LDA 模型都采用 LDA 词袋模式, 得到的词汇表相同, 但词汇重要性排序差异较大。其中 CA-LDA 模型获得的高频共现词汇"中国学术期刊"与交通法学研究主题无关, 主要原因是网络数据中非主题内容词汇的混入。

两种模型前 50 位高频词汇基本都是以"交通运输"、"交通管理"、"交通事故"为主, 核心内容一致。



(a) 交通法学研究高频词汇(533词)共词网络　　(b) 交通法学研究领域高频词汇主题(28个)划分

图 6　交通法学研究高频词汇网络

表 1　传统 LDA 模型与 CA-LDA 模型高频词汇前 50 个词汇对比

模型	高频 50 词(词频降序)
传统 LDA 模型	[1] “交通安全” [2] “交通管理” [3] “交通运输” [4] “交通事故” [5] “交通肇事” [6] “道路交通事故” [7] “交通肇事罪” [8] “交通参与者” [9] “城市轨道交通” [10] “交通警察” [11] “政策法规” [12] “道路交通” [13] “酒后驾驶” [14] “城市交通” [15] “公共交通” [16] “交通运输行业” [17] “强制保险” [18] “节能减排” [19] “道路运输” [20] “交通安全知识” [21] “交强险” [22] “道路交通管理” [23] “水上交通” [24] “交通事故认定” [25] “危险驾驶” [26] “法规司司长” [27] “交通肇事逃逸” [28] “何建中” [29] “交通安全管理” [30] “归责原则” [31] “法律责任” [32] “政策法规司” [33] “交通事故责任” [34] “赔偿责任” [35] “责任认定” [36] “侵权责任法” [37] “安全行车” [38] “司法解释” [39] “逃逸行为” [40] “人民群众” [41] “驾驶经验” [42] “电动自行车” [43] “法律适用” [44] “运输主管部门” [45] “损害赔偿” [46] “交通环境” [47] “责任保险” [48] “公路交通” [49] “新闻发言人” [50] “交通信号”
CA-LDA 模型	[1] “交通安全” [2] “交通管理” [3] “交通肇事” [4] “赔偿责任” [5] “侵权责任法” [6] “归责原则” [7] “交通事故” [8] “道路交通” [9] “酒后驾驶” [10] “道路运输” [11] “强制保险” [12] “《道路交通安全法》” [13] “交通肇事逃逸” [14] “汽车社会” [15] “蓝色经济” [16] “宣传教育” [17] “中国学术期刊” [18] “道路交通管理” [19] “限额范围” [20] “逃逸行为” [21] “结果加重犯” [22] “人身损害赔偿” [23] “交通运输” [24] “机动车安全” [25] “交通事故认定书” [26] “何建中” [27] “交通信号” [28] “电子警察” [29] “机动车” [30] “责任认定” [31] “交通事故责任” [32] “低碳经济” [33] “交通参与者” [34] “《解释》” [35] “政策法规” [36] “政策法规司” [37] “交强险” [38] “交通肇事罪” [39] “交通警察” [40] “交通安全知识” [41] “甩挂运输” [42] “交通安全教育” [43] “损害赔偿” [44] “公路交通” [45] “节能减排” [46] “公交优先” [47] “自由裁量权” [48] “交通事故认定” [49] “法律适用” [50] “责任保险”

比较两种算法结果获得的前 50 位高频词差异:

(1) 两者有 18 个词汇不同(见表 1 中带有底纹词汇);

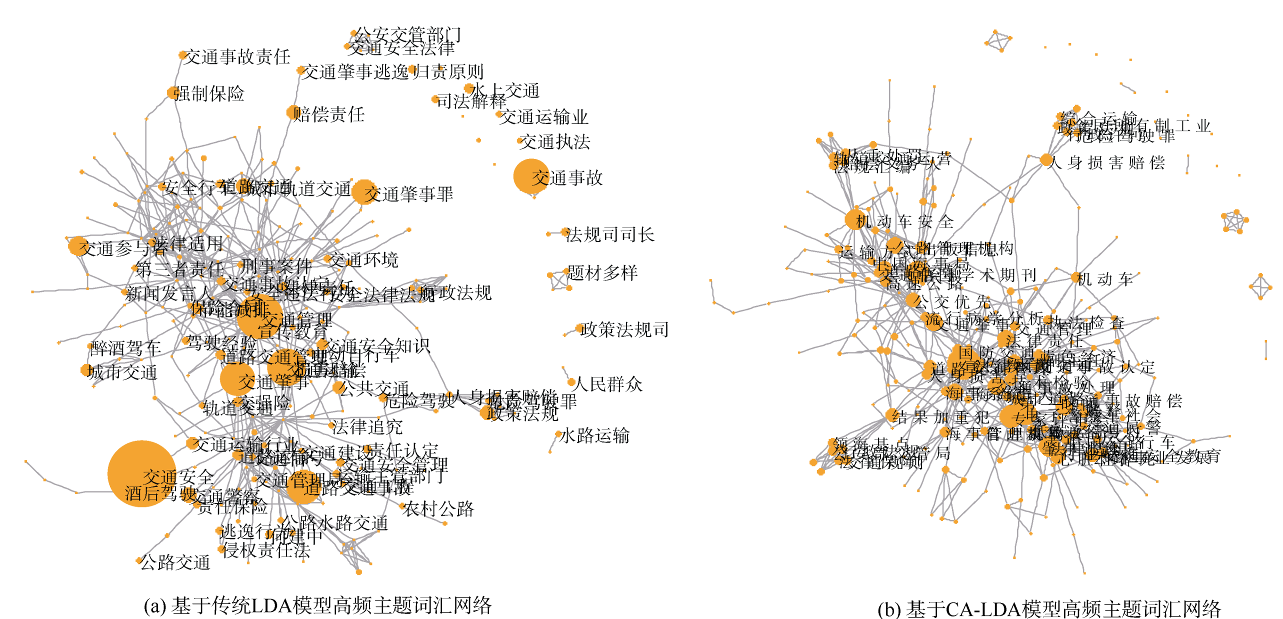
(2) 各词的词频顺序有较大差异;

(3) 传统 LDA 模型生成的主题重点词汇意义较为单一(如“城市轨道交通”、“城市交通”、“公共交通”、“法律责任”、“司法解释”等), CA-LDA 模型结果重点词汇中出现了“汽车社会”、“低碳经济”、“蓝色经济”等研究背景词汇; “《道路交通安全法》”、“《解释》”等法律法规; 以及“限额范围”、“自由裁量权”、“结果加重犯”、“交通事故认定书”、“人身损害赔偿”等争议研究热点内容; “电子警察”、“交通安全教育”、“公交优先”等管理方法。

总之, CA-LDA 模型获得的研究辅助信息比传统 LDA 模型结果要丰富, 而且确实为热点研究内容。

为了显示清晰, 仅对两个模型前 50 位的高频词汇生成词汇网络, 并以节点大小代表词频(或权重修正后词频), 结果如图 7 所示。

从图 7 对比可以看出, 两个模型结果差异较大。



(a) 基于传统LDA模型高频主题词汇网络　　(b) 基于CA-LDA模型高频主题词汇网络

图 7　CA-LDA 模型与传统 LDA 高频主题词汇分布对比

传统LDA模型中孤立的高频词汇较多，说明这些词汇由少量文献产生，而热点应该是多篇文献共同研究内容；传统 LDA 模型生成词频差异较大，分布不均匀，可能把绝对频率较低而相对频率高的词汇作为重点。而 CA-LDA 模型的词频差异较小，关联更强，词汇扎堆明显，主题集聚优势明显。

5　结　语

本文提出一个共词网络分析的 CA-LDA 模型，该模型以网络拓扑结构参数作为主题归类的调节变量，控制词汇主题分配，并使用随机梯度下降技术提高算法执行效率。共词网络拓扑结构参数从词向量关联角度修改词汇分配，其结果不仅反映词频概率，同时，词汇网络的节点介数中心度也能提供信息，从词汇关联上发现枢纽词汇，在纵向上反映领域研究演进的关键技术，在横向上提供解决不同问题的同一有效手段。该模型应用在交通法学研究领域热点主题分析，在处理大规模文献数据中取得了较好效果。相关研究可以拓展应用于各种领域的大规模文献数据自动化处理中。

CA-LDA 模型以节点中心度指标调节 LDA 主题生成，其他复杂网络拓扑结构参数(如节点间紧密度的簇系数、派系、社区)也在不同角度反映共词网络的词汇社交网络关系，可以进一步研究这些参数对LDA模型主题生成的影响；再者，分词是文本分析的重要基础，但是所得结果往往都是单独词汇，存在歧义等特殊性，不利于文本语义分析。本文采用增加合成词的方法来提高语义识别度，这些词汇来自于文献关键词，这种方法不适用于其他文本处理(如网络购物评价等)，可以建立一个领域内的专业词汇表，实现更科学的分词；最后，基于 LDA 主题模型分析需要科学设置主题数 K，虽然该值可以采用混淆度标准交叉验证获得，但在实际分析中计算出的 K 值有时会很大，不利于文献主题的分类整理。未来研究需要找到更为科学的主题数目确定方法，或者对 K 值较大的主题划分结果进一步处理来凝聚主题。

参考文献：

[1]　范云满，马建霞．利用 LDA 的领域新兴主题探测技术综述[J]．现代图书情报技术，2012(12): 58-65. (Fan Yunman, Ma Jianxia. Review on the LDA-based Techniques Detection for the Field Emerging Topic [J]. New Technology of Library and Information Service, 2012(12): 58-65.)

[2]　Day W H E, Edelsbrunner H. Efficient Algorithms for Agglomerative Hierarchical Clustering Methods [J]. Journal of Classification, 1984, 1(1): 7-24.

[3]　曹高辉，焦玉英，成全．基于凝聚式层次聚类算法的标签聚类研究[J]．现代图书情报技术，2008(4): 23-28. (Cao Gaohui, Jiao Yuying, Cheng Quan. Research on Tag Cluster

Based on Hierarchical Agglomerative Clustering Algorithm [J]. New Technology of Library and Information Service, 2008(4): 23-28.)

[4] Katz S. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer [J]. IEEE Transactions on Acoustics, Speech, & Signal Processing, 1987, 35(3): 400-401.

[5] 陈浪舟，黄泰翼. 一种新颖的词聚类算法和可变长统计语言模型[J]. 计算机学报，1999，22(9): 942-948. (Chen Langzhou, Huang Taiyi. A Novel Word Clustering Algorithm and Vari-Gram Language Model [J]. Chinese Journal of Computers, 1999, 22(9): 942-948.)

[6] Salton G, Wong A, Yang C S. A Vector Space Model for Automatic Indexing [J]. Communications of the ACM, 1975, 18(11): 613-620.

[7] 庞剑锋，卜东波，白硕. 基于向量空间模型的文本自动分类系统的研究与实现[J]. 计算机应用研究，2001，27(9): 23-26. (Pang Jianfeng, Bu Dongbo, Bai Shuo. Research and Implementation of Text Categorization System Based on VSM [J]. Application Research of Computers, 2001, 27(9): 23-26.)

[8] Porteous I, Newman D, Ihler A, et al. Fast Collapsed Gibbs Sampling for Latent Dirichlet Allocation [C]. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, 2008: 569-577.

[9] Newman D, Asuncion A, Smyth P, et al. Distributed Inference for Latent Dirichlet Allocation [C]. In: Proceedings of the 21st Annual Conference on Neural Information Processing Systems. 2007: 1081-1088.

[10] Asuncion A U,Smyth P, Welling M. Asynchronous Distributed Learning of Topic Models [C]. In: Proceedings of the 22nd Annual Conference on Neural Information Processing Systems.2008: 81-88.

[11] Blei D M, Lafferty J D. A Correlated Topic Model of Science [J]. The Annals of Applied Statistics, 2007, 1(1): 17-35.

[12] Sato I, Nakagawa H. Topic Models with Power-law Using Pitman-Yor Process [C]. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, 2010: 673-682.

[13] Teh Y W. Dirichlet Process [A]. //Sammut C, Webb G I. Encyclopedia of Machine Learning [M]. Springer US, 2011: 280-287.

[14] Callon M, Courtial J P, Turner W, et al. From Translations to Problematic Networks: An Introduction to Co-word Analysis [J]. Social Science Information, 1983, 22(2): 191-235.

[15] Callon M, Courtial J P, Laville F. Co-word Analysis as a Tool for Describing the Network of Interactions Between Basic and Technological Research: The Case of Polymer Chemsitry [J]. Scientometrics, 1991, 22(1): 155-205.

[16] Coulter N, Monarch I, Konda S. Software Engineering as Seen Through Its Research Literature: A Study in Co-word Analysis [J]. Journal of the American Society for Information Science, 1998, 49(13): 1206-1223.

[17] 张晓冬，周宏丽，胡杨,等. 基于共词分析和社会网络分析的我国计算机集成制造系统研究热点[J]. 科技管理研究，2016(11): 145-149. (Zhang Xiaodong, Zhou Hongli, Hu Yang, et al. Research Hotspots of Computer Integrated Manufacturing of China Based on Co-word Analysis and Social Network Analysis [J]. Science and Technology Management Research, 2016(11): 145-149.)

[18] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[19] Newman D, Bonilla E V, Buntine W. Improving Topic Coherence with Regularized Topic Models [C]. In: Proceedings of the 24th International Conference on Neural Information Processing Systems.2011: 496-504.

[20] Jordan M I, Ghahramani Z, Jaakkola T S, et al. An Introduction to Variational Methods for Graphical Models [J]. Machine Learning, 1999, 37(2): 183-233.

[21] Hoffman M, Blei D, Wang C, et al. Stochastic Variational Inference [J]. Journal of Machine Learning Research, 2013, 14(1): 1303-1347.

[22] Brandes U. A Faster Algorithm for Betweenness Centrality [J]. Journal of Mathematical Sociology, 2001, 25(2): 163-177.

[23] Newman M E J. The Structure and Function of Complex Networks [J]. SIAM Review, 2003, 45(2): 167-256.

作者贡献声明:

马红：提出研究思路，设计研究方案，结论分析；

蔡永明：数据获取、整理和清洗，算法设计，程序开发。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据见期刊网络版 http://www.infotech.ac.cn。

[1] 马红, 蔡永明. CNKI 检索原始数据.rar. CNKI 检索原始数据.

[2] 马红, 蔡永明. 合成词库和停用词库.rar. 合成词库和停用词库.

[3] 马红, 蔡永明. 描述性统计分析数据.xlsx. 描述性统计分析数据.

[4] 马红, 蔡永明. 预处理后数据.rar. 预处理后数据.

收稿日期: 2016-08-01
收修改稿日期: 2016-11-02

A CA-LDA Model for Chinese Topic Analysis: Case Study of Transportation Law Literature

Ma Hong[1]　Cai Yongming[2]
[1] (School of Transportation Law, Shandong Jiaotong University, Jinan 250357, China)
[2] (Business School, University of Jinan, Jinan 250022, China)

Abstract: [**Objective**] This paper aims to improve the effectiveness of extracting Chinese literature topics with the help of LDA model and co-word network analysis. [**Methods**] First, we added keywords to the word segmentation dictionary for the abstracts, which improved the semantic recognition of topic analysis. Second, we proposed a Latent Dirichlet Allocation Model with Co-word Analysis (CA-LDA) to control the topic distribution generated by the weight of co-word network topology parameters (i.e. Betweenness Centrality). Finally, we extracted the words with high connectivity (Betweenness Centrality) and frequency. [**Results**] The CA-LDA model retrieved high frequency and high connectivity words simultaneously, which were important for subject analysis. The proposed algorithm could also identify key node technical vocabularies with the help of co-word analysis. [**Limitations**] The K value (number of topics) was obtained by cross validation with perplexity. Thus, it was difficult to classify the document topics with larger K value. More research is needed to deal with this issue. [**Conclusions**] The proposed model effectively analyzes the topics of Chinese literature on transportation laws, which could also process literature data from other fields automatically.

Keywords: Latent Dirichlet Allocation Model with Co-word Analysis　Co-words　Network topology parameters　Stochastic gradient descent　Key word in transportation law literature